ARGOS: ADAPT RICH GEOMETRIC PRIORS FOR GENERALIZABLE ONLINE SCENE-CHANGE-DETECTION

Anonymous authors
Paper under double-blind review

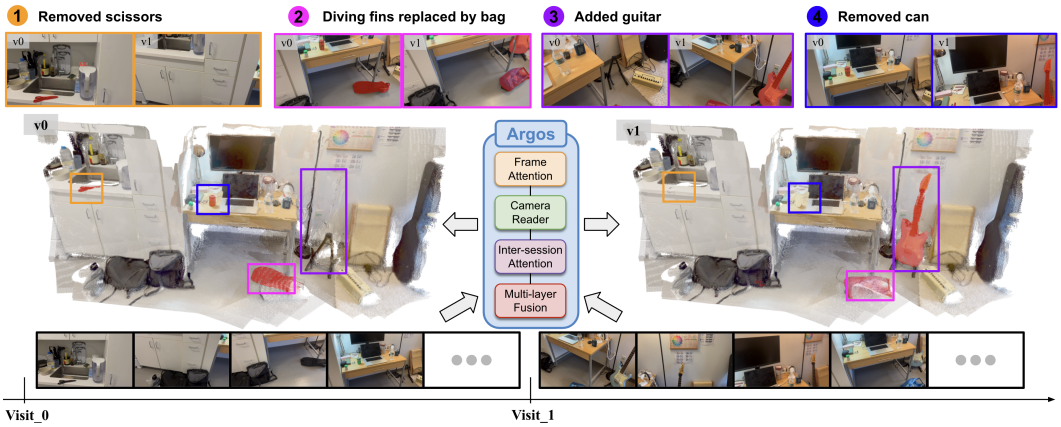


Figure 1: We propose **Argos**, a feedforward model for sub-second change detection and 3D reconstruction, and integrate it into Argos-SLAM for online 4D reconstruction of change-aware point-cloud maps. The figure illustrates how Argos captures scene changes, including the added guitar (3) and removed can (4), while maintaining a consistent spatio-temporal representation across visits.

ABSTRACT

Robots operating in dynamic environments require reliable detection of how their surroundings change over time. Existing learning-based methods largely rely on pairwise 2D image features, which struggle under large viewpoint changes and occlusions, are sensitive to noise, and show limited generalization across domains, while explicit 3D approaches typically require costly offline optimization. We show that the implicit 3D knowledge of Geometric Foundation Models (GFMs) provides a strong basis for addressing these limitations. We introduce Argos, which adapts GFM features for joint scene change detection and 3D reconstruction. To address data scarcity and take a step toward a foundation model for scene change detection, we introduce a large-scale benchmark comprising two synthetic datasets and one real-world dataset, and train jointly across diverse datasets to improve cross-domain generalization. We further introduce Argos-SLAM, a real-time system designed for robotics, which performs online change detection and change-aware 4D mapping. Across benchmarks, our framework substantially outperforms existing baselines, with gains of up to 42.01% in change IoU and 27.91% in F1, while supporting scalable deployment in changing real-world environments. Code and datasets are available at `coming_soon`.

1 INTRODUCTION

Long-term autonomy requires agents to continuously update their spatial knowledge in the face of a changing world. A moved tool may invalidate a planned interaction, while newly placed furniture may block a previously traversable route. For robots to operate autonomously in such environments, they must not only perceive and understand how the environment changes over time, but also maintain an accurate memory of those changes for subsequent navigation and manipulation. A map that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

faithfully reflects the current state of the environment is therefore a key component of long-term robot autonomy. This motivates the need for robust change detection that can operate under large viewpoint variations, occlusions, and continuous observations during real-world robot operation.

Existing approaches have progressed from image pairs to multi-view observations, but geometric ambiguity and limited generalization remain central challenges. Learned detectors predominantly compare 2D features, making large viewpoint shifts and occlusions difficult to resolve (Wang et al., 2023; Lin et al., 2025; Yoon & Kim, 2026b;a). Explicit 3D pipelines provide spatial context, but often infer change through heavy optimization with handcrafted comparison rules (Wu et al., 2025; Galappaththige et al., 2026). Meanwhile, benchmark-specific training and scarce long-sequence annotations restrict generalization across environments and observation trajectories (Kim & Kim, 2025; Yoon & Kim, 2026a).

To address these limitations, we introduce **Argos** (**A**dapt **R**ich **G**eometric Priors for Generalizable **O**nline **S**cene-Change-Detection), a feedforward model that learns changes directly from the implicit representations of Geometric Foundation Models (GFMs) (Wang et al., 2025; 2026). Our key insight is that the geometric knowledge learned to associate scene content across viewpoints also provides a prior for recognizing inconsistent content across sessions. Crucially, such inconsistency requires to be interpreted through co-visibility: observing that the space previously occupied by an object is now empty provides evidence of removal, whereas simply failing to see an occluded object does not. We realize this insight through a co-visibility-aware inter-session comparator that incorporates camera information and session membership to interpret feature differences. As illustrated in Figure 1, Argos jointly predicts change masks and scene geometry from two unposed RGB image collections, retaining reconstruction capability while learning to compare different scene states.

To support learning beyond individual benchmarks, we introduce Argos-CD, which comprises two large-scale synthetic datasets with automatically generated robot traversals and a challenging real-world evaluation set. Joint training on Argos-CD and existing datasets provides diverse change supervision across domains, taking a step toward a foundation model for scene change detection. Experiments demonstrate substantial gains across paired-image and video benchmarks; even with synthetic-only change supervision, Argos surpasses learned baselines fine-tuned on real data on real-world evaluations. We further integrate Argos with VGGT-SLAM (Maggio & Carlone, 2026) into Argos-SLAM. This system supports online reconstruction and change updates for change-aware 4D mapping.

Our contributions are summarized as follows:

- **Geometrically grounded change detection.** We introduce Argos, a model that adapts implicit 3D priors through co-visibility-aware inter-session comparison for joint change detection and reconstruction from a pair of RGB sequences.
- **Large-scale dataset and generalization.** We introduce the Argos-CD dataset and demonstrate that leveraging pretrained GFM features together with multi-dataset training enables strong benchmark performance and synthetic-to-real transfer.
- **Online change-aware mapping.** We develop Argos-SLAM, which connects learned change predictions to online reconstruction of a change-aware point-cloud map.

2 Related Work

Geometric Foundation Model. Multi-view geometry and SLAM establish the spatial relationships needed to reconstruct scenes from visual observations (Hartley & Zisserman, 2003; Cadena et al., 2016). Geometric Foundation Models (GFMs), including VGGT and VGGT-Ω, predict camera poses and dense geometry through feedforward inference (Wang et al., 2025; 2026). D4RT and PAGE-4D extend feedforward geometric perception to dynamic scenes (Zhang et al., 2026a; Zhou et al., 2026), while WT-DiT connects spatial reconstruction with generative world modeling (Zhang et al., 2026b). Beyond reconstruction, Bratulić et al. (2026) probe VGGT's internal representations to investigate the geometric information encoded by GFMs, revealing that these representations implicitly capture epipolar geometry. We study how these reconstruction-trained features can support co-visibility-aware comparison of different scene states across sessions. Rather than using only reconstruction outputs, our approach uses GFM's latent representations and tokens to inform change detection.

Change Detection. Scene change detection (SCD) traditionally localizes changes between image pairs from similar viewpoints (Sakurada & Okatani, 2015; Alcantarilla et al., 2018; Sakurada et al., 2020; Park et al., 2021), using learned pairwise features, temporal interaction, segmentation, or cross-view matching (Varghese et al., 2018; Caye Daudt et al., 2018; Chen et al., 2021; Wang et al., 2023; Sachdeva & Zisserman, 2023b;a). Recent methods leverage pretrained representations such as DINOv2 (Oquab et al., 2023) and SAM (Kirillov et al., 2023) through task-specific adaptation (Lin et al., 2025; Yoon & Kim, 2026b) or zero-shot comparison (Cho et al., 2025; Kannan & Min, 2025; Kim & Kim, 2025). Multi-view work extends SCD to unaligned observations (Wald et al., 2019; Wu et al., 2025; Galappaththige et al., 2025): VSCD (Yoon & Kim, 2026a) learns multi-reference cross-frame correspondence, SceneDiff (Wu et al., 2025) selects co-visible views using reconstructed geometry, and O-SCD (Galappaththige et al., 2026) performs self-supervised fusion over an offline Gaussian Splatting reconstruction (Kerbl et al., 2023). However, learned methods can still struggle with cross-domain transfer, while zero-shot approaches often require costly reconstruction or optimization. In contrast, Argos performs cross-session comparison directly in the latent space of a frozen GFM and improves transfer through joint training across diverse datasets.

Robot Mapping in Dynamic Environments. The literature on long-term mapping and SLAM has investigated robot operation in changing environments (Fehr et al., 2017; Fu et al., 2022; Schmid et al., 2022). Khronos (Schmid et al., 2024) constructs dense spatio-temporal maps, identifying quasi-static changes through global optimization over TSDF-based maps (Curless & Levoy, 1996). SuperMap (Zhao et al., 2026) extends this direction with a real-time online system and open-set semantic capabilities using vision foundation models (Oquab et al., 2023; Ravi et al., 2025). For LiDAR, Chamelion (Jang et al., 2026) uses a dual-head network for change classification and long-term map maintenance, while ELite (Gil et al., 2025) models ephemerality for lifelong mapping. These systems typically detect change through explicit map geometry or object associations rather than leveraging advances in learned change detection. We bridge the two with Argos-SLAM, which combines online reconstruction and change detection to maintain a change-aware point-cloud map: SLAM supplies the spatio-temporal structure for long-term deployment, while learned RGB change detection improves change identification beyond geometry alone.

3 Problem Formulation

We consider a robot that revisits a scene and needs to detect changes while reconstructing the scene.

Two-visit Change Detection. The inputs are two sets of *unposed* RGB images from two visits, $\mathbb{I}^{v_0} = \{I_i^{v_0}\}_{i=1}^{K}$ and $\mathbb{I}^{v_1} = \{I_j^{v_1}\}_{j=1}^{N}$. The visits may differ in length, trajectory, viewpoint, and overlap. The goal is to predict a binary change mask for every image,

$$\mathbb{M}^{v_0} = \{M_i^{v_0}\}_{i=1}^{K}, \qquad \mathbb{M}^{v_1} = \{M_j^{v_1}\}_{j=1}^{N}, \qquad M \in \{0,1\}^{H\times W}. \tag{1}$$

Changed content is labeled in the visit where it is physically present. For instance, if a new object becomes visible in $I_j^{v_1}$, the corresponding pixels are labeled in $M_j^{v_1}$. Conversely, if an object visible in $I_i^{v_0}$ is absent from a co-visible frame $I_j^{v_1}$, its pixels are labeled in $M_i^{v_0}$.

Change-aware Point-cloud Map. Naively reconstructing the scene from both visits produces a joint point cloud $\mathbb{P}$ that mixes the static background with changed elements observed across visits. Instead, we seek a change-aware point-cloud map $\mathbb{P}(v)$, where $v \in \{v_0, v_1\}$ specifies the queried visit, that disentangles these changes and faithfully represents the corresponding scene state. Specifically, $\mathbb{P}(v_0)$ removes geometry corresponding to elements present only in v_1, while retaining the shared static background and visit-specific geometry from v_0; $\mathbb{P}(v_1)$ is defined symmetrically. For example, if a cup in v_0 is replaced by a bowl in v_1, the joint reconstruction may contain overlapping geometry from both objects. Querying v_0 removes the bowl geometry and recovers the scene with the cup.

4 Method

This section describes Argos. We first motivate our design, then describe the architecture, training data, and finally integration with a state-of-the-art SLAM system to form Argos-SLAM.

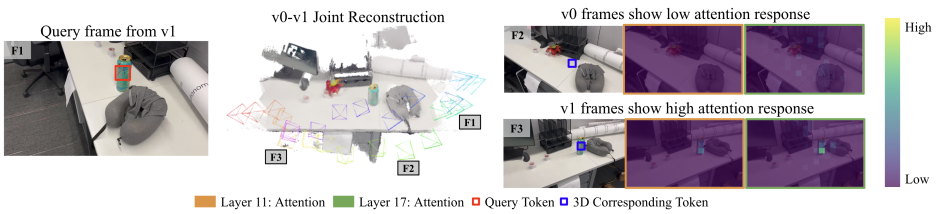


Figure 2: **Attention score visualization of selected layers.** The token associated with the added green bottle in v_1 shows strong attention to the corresponding bottle region across different views within the same visit, while finding no high-attention score in v_0, where the bottle is absent.

4.1 Motivation

Learning-based scene change detection often compares 2D image features, making it difficult to distinguish true changes from viewpoint- or occlusion-induced appearance differences. We instead view change detection as inherently 3D: evidence for change requires both observing the same scene region across visits and finding inconsistent content there.

GFMs provide a natural representation for this reasoning. In particular, VGGT-Ω (Wang et al., 2026) is pretrained with a matching loss $\mathcal{L}_{\text{match}}$ that encourages features corresponding to the same 3D location to align across views. Inspired by the attention analysis used in PAGE-4D (Zhou et al., 2026) and VGGT-SLAM 2.0 (Maggio & Carlone, 2026), we similarly probe pretrained VGGT-Ω across visits. Figure 2 shows that its intermediate features respond strongly to corresponding object locations across viewpoints, but weakly when the object is absent, suggesting useful cues for co-visibility and cross-visit consistency. Guided by this observation, Argos uses pose and session information to adapt these implicit geometric features for co-visibility-aware cross-session comparison, jointly predicting scene changes, camera poses, and depth.

4.2 Model Architecture

Illustrated in Figure 3, Argos is a unified model that jointly predicts change masks and scene geometry from the two visits. To enable cross-visit change reasoning, it augments VGGT-Ω with a *Co-visibility-aware Spatio-Temporal Comparator* comprising a Camera Reader, Inter-session Cross-Attention, and a DPT Fusion Head. We index the combined $n = K + N$ images by $r \in \{1, \dots, n\}$, with the first K images from v_0 and the remaining N from v_1. The model implements

$$f_\theta\big(\mathbb{I}^{v_0}, \mathbb{I}^{v_1}\big) = \big\{\hat{M}_r, \hat{D}_r, \hat{g}_r\big\}_{r=1}^{n}, \tag{2}$$

where $\hat{M}_r \in \{0,1\}^{H\times W}$ is a change mask, $\hat{D}_r \in \mathbb{R}^{H\times W}$ is a depth map, and $\hat{g}_r \in \mathbb{R}^9$ encodes camera extrinsics and intrinsics. Unprojecting the predicted depth using the estimated cameras and disentangle changed regions with the masks yields a change-aware point-cloud map $\mathbb{P}(v)$.

Backbone and Tokens. The frozen VGGT-Ω backbone (Wang et al., 2026) processes both visits jointly. We extract its features at layers $\mathcal{L} = \{4, 11, 17, 23\}$ and denote the patch tokens for image r at layer ℓ by $X_r^\ell \in \mathbb{R}^{T\times d}$, where T is the number of image patches and $d = 2048$ is the feature width. The pretrained camera and depth heads read the original backbone features; the following modules adapt them for the change branch.

Camera Reader. To make patch tokens co-visibility aware, inspired by DPT's readout operator (Ranftl et al., 2021), Camera Reader injects final camera features into patch tokens after selected VGGT-Ω intermediate layers. Let $c_r \in \mathbb{R}^d$ denote the normalized camera-head token for image r, immediately before it is decoded into $\hat{g}_r$. For each patch $q \in \{1, \dots, T\}$ and layer $\ell \in \mathcal{L}$, we compute

$$\tilde{x}_{r,q}^\ell = x_{r,q}^\ell + \mathrm{MLP}_\ell\big([c_r\,;\, x_{r,q}^\ell]\big), \tag{3}$$

where $x_{r,q}^\ell$ is the q-th row of X_r^ℓ, $[\cdot\,;\,\cdot]$ denotes channel-wise concatenation, and $\mathrm{MLP}_\ell : \mathbb{R}^{2d} \to \mathbb{R}^d$ is learned separately for each layer. The same c_r is broadcast to every patch of image r at all selected layers. MLP implementation details are given in Appendix A.1.

Inter-session Cross-Attention. The backbone mixes information across all images without explicitly distinguishing the two visits. We therefore use visit membership to compare their features at

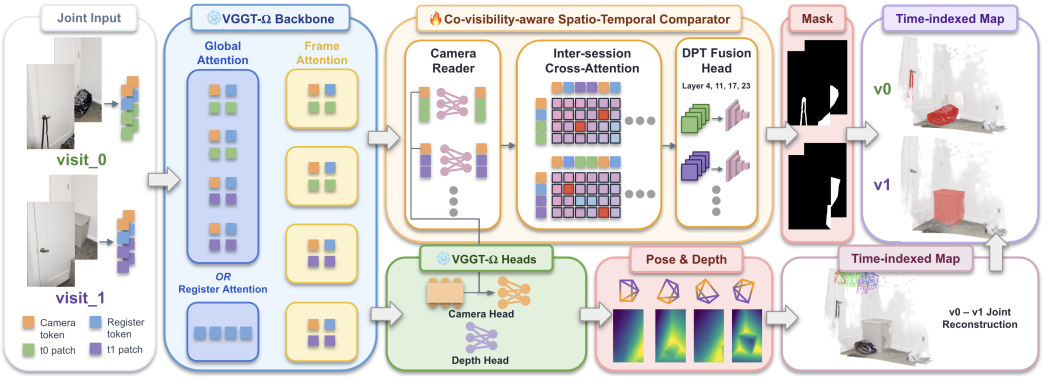


Figure 3: **Argos Architecture.** Building on VGGT-Ω, we introduce a Co-visibility-aware Spatio-Temporal Comparator with three modules: a Camera Reader that injects final camera tokens into intermediate patch features, Inter-session Cross-Attention that aggregates information across visits, and a lightweight DPT Fusion Head that combines multi-layer features to predict per-visit change masks. These masks are then used to disentangle the joint reconstruction into a change-aware map.

each layer $\ell \in \mathcal{L}$. Let $Z^{\ell}_{v_0}$ and $Z^{\ell}_{v_1}$ stack the camera, register, and camera-aware patch tokens from their respective visits. Each visit uses its own tokens as queries and the other visit's tokens as context, giving the residual updates

$$\bar{Z}^{\ell}_{v_0} = Z^{\ell}_{v_0} + \mathrm{CrossAttn}_{\ell}\big(Z^{\ell}_{v_0}, Z^{\ell}_{v_1}\big), \qquad \bar{Z}^{\ell}_{v_1} = Z^{\ell}_{v_1} + \mathrm{CrossAttn}_{\ell}\big(Z^{\ell}_{v_1}, Z^{\ell}_{v_0}\big), \tag{4}$$

where $\mathrm{CrossAttn}_{\ell}(A, B)$ denotes multi-head attention with queries from A and keys and values from B, following the standard attention formulation (Vaswani et al., 2017). A feed-forward sub-layer then maps each $\bar{Z}^{\ell}_{v}$ to $\tilde{Z}^{\ell}_{v}$, the features passed to the change head. Since change signal is symmetric, the two directions share attention and feed-forward weights, while each feature layer has its own block. The full block equations are given in Appendix A.1.

Intuitively, camera-augmented attention compares both co-visibility and visual appearance. Together, these learned interactions connect the co-visibility and visual inconsistency intuition in Section 4.1, enabling the change head to interpret appearance differences in their geometric context.

DPT Fusion Head. We extract the updated patch tokens $Y^{\ell}_{r} \in \mathbb{R}^{T \times d}$ for image r from $\tilde{Z}^{\ell}_{v}$ and decode them with a DPT-style head $\mathcal{H}_{\mathrm{CD}}$ (Ranftl et al., 2021), following VGGT-Ω's depth decoder. The head reshapes the tokens into spatial feature maps, resizes and fuses them across layers, and predicts per-pixel logits $z_r \in \mathbb{R}^{H \times W}$:

$$z_r = \mathcal{H}_{\mathrm{CD}}\big(\{Y^{\ell}_{r}\}_{\ell \in \mathcal{L}}\big), \qquad \hat{p}_r = \sigma(z_r), \qquad \hat{M}_r = \mathbf{1}[\hat{p}_r > 0.5], \tag{5}$$

where σ is the sigmoid function and the threshold is applied independently to each pixel.

4.3 New Dataset

Video change-detection data remain scarce for long robot traversals, hindering large-scale training. Therefore, we introduce **Argos-CD**, a large-scale benchmark comprising two synthetic datasets, **Argos-CD-HSSD** and **Argos-CD-SceneSmith**, leveraging simulation assets from HSSD (Khanna et al., 2024) and SceneSmith (Pfaff et al., 2026), and a real-world evaluation set **Argos-CD-real**. Our synthetic datasets share an automated pipeline that samples scene states, plans independent trajectories to cover scene objects and navigable areas, and generates change masks from object identities. Table 1 shows that Argos-CD-HSSD covers $4\times$ more visually distinct places on average and up to $2\times$ longer maximum video duration than VSCD and SceneDiff. Furthermore, **Argos-CD-real** captures challenging real-world scenarios, including cluttered workshops and opposite-direction camera traversals. See Appendix A.2 for additional details.

4.4 Training

Training Datasets. We train on VL-CMU-CD (Alcantarilla et al., 2018), PSCD (Sakurada et al., 2020), and ChangeSim (Park et al., 2021) from the SCD task, alongside VSCD (Yoon & Kim,

Table 1: **Dataset comparison.** Places* and single-visit video length (seconds) are reported as mean / min / max. Pairs counts cross-visit comparisons and Views counts frames with change masks. *Approximated through appearance clustering.

	Dataset	Pairs	Views	Scenes	Places*	Length (s)	Modality	Viewpoint	Trajectory
Released datasets	VL-CMU-CD	1,362	1,362	152	–	–	Real	Identical	-
	PSCD	770	1,540	–	–	–	Real	Identical	-
	ChangeSim	20	130 K	10	–	–	Sim.	Similar	Manual
	VSCD	1,041	1.13 M	210	1.6/1/7	36.0/10.1/116.0	both	Free	Manual
	SceneDiff Dataset	355	172 K	355	1.1/1/2	8.1/2.4/26.1	Real	Free	Manual
Argos-CD (Ours)	Argos-CD-HSSD	501	644 K	167	**3.8/1/11**	**42.8/4.0/203.6**	Sim.	Free	Automatic
	Argos-CD-SceneSmith	612	138 K	204	1.5/1/5	7.5/1.2/51.7	Sim.	Free	Automatic
	Argos-CD-real	8	5,514	8	1.0/1/1	11.5/4.8/21.7	Real	Free	Manual

2026a), Argos-CD-HSSD, Argos-CD-SceneSmith, and SceneDiff (Wu et al., 2025) from the VSCD task. We evaluate on the held-out test sets of each benchmark, as well as on Argos-CD-real for additional out-of-distribution real-world evaluation.

Loss Function and Training Details. To handle class imbalance between sparse change pixels and dominant static regions, we apply weighted binary cross-entropy exclusively to valid (labeled) pixels. We freeze the original VGGT-Ω components and train only the three modules constituting the Co-visibility-aware Spatio-Temporal Comparator. Furthermore, we extend VGGT's training recipe by interleaving SCD and VSCD datasets across input sequence lengths ranging from 2 to 64 frames. See Appendix A.1 for additional implementation details.

4.5 Online Robot Mapping System

Online Reconstruction, Asynchronous Change-aware Map Updates. We introduce Argos-SLAM, an extension of VGGT-SLAM 2.0 (Maggio & Carlone, 2026) for online change detection and change-aware 4D mapping. The same Argos network serves as both (i) the *geometry* backbone for submap reconstruction and (ii) the *change* model for cross-visit comparison. Given a prior reconstruction from the first visit v_0, the system incrementally reconstructs the scene during the second visit v_1 and aligns the incoming geometry with v_0 in a shared coordinate frame. Each new submap is integrated into the global reconstruction and triggers an asynchronous change-detection job. Once completed, the detected changes are incorporated into the change-aware point-cloud map $\mathbb{P}(v)$, which therefore reflects all geometry and changes accumulated up to that submap. To maintain online operation, each job retrieves only the most relevant prior keyframes rather than comparing against the entire prior visit. An efficient retrieval cascade combines cached SALAD descriptor matching, strided geometric reprojection filtering, and scene-centroid coverage search, while capping the combined input at 64 images for sub-second inference. Change-detection jobs run asynchronously alongside reconstruction, allowing the change-aware map to be updated incrementally as new submaps arrive. Section 5.6 reports the runtime during real-world robot deployment, while Appendix A.3 provides additional implementation details.

5 Experiments

5.1 Experimental Setup

Baselines and Benchmark. We evaluate Argos on the real and synthetic SCD and VSCD datasets described in Section 4.4. We compare against learned SCD methods (FC-EF, FC-Siam-diff, FC-Siam-conc (Caye Daudt et al., 2018), CSCDNet (Sakurada et al., 2020), DR-TANet (Chen et al., 2021), C-3PO (Wang et al., 2023), RobustSCD (Lin et al., 2025), TERDNet (Yoon & Kim, 2026b)), learned VSCD (VSCDNet (Yoon & Kim, 2026a)), zero-shot SCD (ZSSCD (Cho et al., 2025), GeSCF (Kim & Kim, 2025)) and VSCD (SceneDiff (Wu et al., 2025)) methods. We additionally evaluate the zero-shot Gaussian-splatting method O-SCD (Galappaththige et al., 2026); due to frequent failures, its results are deferred to Appendix A.5. Appendix A.4 shows more details.

Model Variants. We consider four variants: **Ours**, trained specifically on each evaluation dataset; **Ours-synthetic-only**, trained exclusively on synthetic data; **Ours-finetune**, obtained by fine-tuning the synthetic variant on real-world data; and **Ours-foundation**, trained on all available datasets.

Table 2: **Real-world generalization study.** Argos performs consistently better than learned baselines even without fine-tuning to the real-world datasets. † indicates zero-shot methods.

	Method	SceneDiff Dataset				Argos-CD-real			
		Static	Change	mIoU	F1	Static	Change	mIoU	F1
SCD	CSCDNet-finetune	93.33	15.25	54.29	20.23	94.75	8.88	51.81	14.20
	C-3PO* -finetune	94.42	17.88	56.15	22.47	95.49	8.71	52.10	12.41
	GeSCF†	69.69	5.04	37.37	15.12	63.64	2.95	33.30	11.13
	RobustSCD-finetune	94.41	18.97	56.69	21.25	94.46	9.06	51.76	15.92
	ZSSCD†	88.63	6.74	47.69	13.22	79.97	3.79	41.88	11.47
	TERDNet-finetune	94.30	20.32	57.30	25.14	93.90	8.56	51.23	16.75
VSCD	VSCDNet	91.40	10.48	50.94	15.46	91.06	7.26	49.16	12.91
	VSCDNet-finetune	95.72	20.70	58.21	23.84	96.76	11.72	54.24	17.88
	VSCDNet-foundation	95.85	19.27	57.56	24.26	93.75	9.54	51.65	19.77
	SceneDiff†	96.48	30.31	63.40	37.58	95.50	19.49	57.50	36.90
	Ours-synthetic-only	97.13	39.84	68.48	37.42	95.88	20.87	58.37	34.17
	Ours-finetune	98.62	59.91	79.27	44.73	**97.69**	**36.29**	**66.99**	39.38
	Ours-foundation	**98.78**	**62.54**	**80.66**	**45.98**	97.36	29.50	63.43	**39.90**

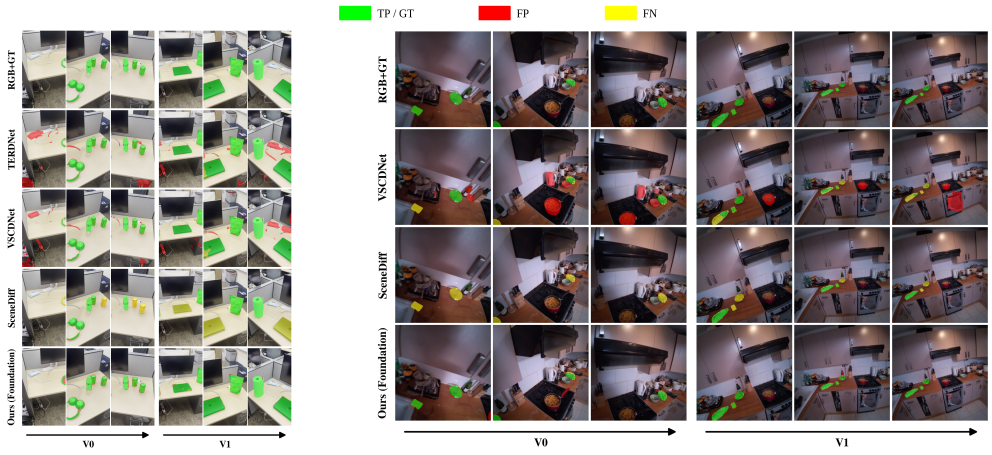


Figure 4: **Qualitative results.** We show results of the best-performing models. Argos shows cleaner masks and fewer false positives or false negatives in real-world. Appendix A.7 shows more results.

Metrics. We treat change detection as binary segmentation and report static IoU, change IoU, their mean (mIoU), and change-class macro F1, all as percentages.

5.2 Real-World Generalization

Table 2 shows that **Ours-synthetic-only** substantially outperforms the best fine-tuned learned baselines in change IoU (39.84 vs. 20.70 on SceneDiff and 20.87 vs. 11.72 on Argos-CD-real) and also surpasses the zero-shot SceneDiff baseline. We attribute this sim-to-real generalization to the GFM's implicit 3D prior and generalization ability, which relies less on domain-specific appearance than 2D baselines. Real-world fine-tuning further improves performance, while **Ours-foundation** matches or exceeds **Ours-finetune**, suggesting that broader training can enable near out-of-the-box deployment. Training the closest baseline, VSCDNet, with the same strategy as Ours-foundation still yields substantially lower performance (19.27 vs. 62.54 change IoU on SceneDiff), supporting the importance of our architecture rather than training data alone. Figure 4 shows qualitative results.

5.3 Video Scene Change Detection (VSCD)

Table 3 shows that Argos substantially outperforms all baselines in all metrics across all benchmark, while **Ours-foundation** retains a similarly strong advantage. Particularly, on Argos-CD-SceneSmith, our model increase 42.01% on change IoU and 27.91% on F1. These benchmarks are challenging due to large viewpoint changes and occlusions, where parallax and disocclusion can be mistaken for scene changes even after frame matching. By leveraging co-visibility-aware cross-

Table 3: **Performance on Video Scene Change Detection.** *C-3PO architecture disables the "remove" feature branch. **The VSCD dataset is modified to contain "appeared" labels in both reference and query frame.

Method	Argos-CD-HSSD				Argos-CD-SceneSmith				VSCD**			
	Static	Change	mIoU	F1	Static	Change	mIoU	F1	Static	Change	mIoU	F1
CSCDNet	92.91	38.10	65.51	36.53	94.32	31.11	62.72	34.83	94.03	31.95	62.99	32.44
C-3PO*	91.85	37.65	64.75	37.68	94.46	38.68	66.57	40.77	93.88	32.98	63.43	32.90
GeSCF†	76.73	13.81	45.27	28.41	81.56	13.72	47.64	29.80	77.74	12.26	45.00	23.98
RobustSCD	88.86	33.18	61.02	34.76	92.70	33.92	63.30	36.85	89.54	27.18	58.36	30.59
ZSSCD†	90.03	16.58	53.31	16.22	91.63	15.15	53.39	15.49	93.00	15.72	54.36	16.02
TERDNet	93.05	42.14	67.59	40.62	95.19	38.58	66.88	40.31	90.90	27.92	59.41	30.92
VSCDNet	93.50	42.71	68.11	41.77	94.62	38.94	66.78	41.17	94.49	34.55	64.52	36.31
SceneDiff†	93.80	33.44	63.62	32.44	96.38	40.42	68.40	36.58	96.71	48.21	72.46	38.70
Ours	**98.38**	**78.64**	**88.51**	**62.16**	**99.08**	**82.43**	**90.75**	**69.08**	**98.24**	**68.69**	**83.46**	53.14
Ours-foundation	98.07	75.51	86.79	60.69	98.91	80.41	89.66	68.93	98.05	68.52	83.28	**53.33**

Table 4: **Performance on paired-image Scene Change Detection.** Argos outperforms all baseline methods, and the foundation variant of Argos scores closely to the dataset specific variants.

Method	PSCD				ChangeSim*				VL-CMU-CD			
	Static	Change	mIoU	F1	Static	Change	mIoU	F1	Static	Change	mIoU	F1
FC-EF	84.97	37.99	61.48	48.14	66.21	12.44	39.32	20.72	91.36	36.72	64.04	45.61
FC-Siam-diff	89.55	43.72	66.64	52.67	85.77	21.12	53.44	33.61	96.17	55.97	76.07	64.21
FC-Siam-conc	86.96	41.76	64.36	52.13	81.63	19.01	50.32	29.40	94.40	50.81	72.61	58.71*
CSCDNet	92.57	48.84	70.70	60.81	92.77	26.48	59.62	37.64	97.57	69.17	83.37	76.90
DR-TANet	91.17	43.40	67.28	55.79	90.91	27.57	59.24	38.53	97.16	65.81	81.48	73.98
C-3PO*	92.36	50.19	71.27	60.66	93.00	28.68	60.84	38.92	97.86	72.33	85.10	80.00
ZSSCD†	83.41	19.94	51.67	28.94	94.55	30.03	62.29	40.51	93.12	36.14	64.63	50.14
RobustSCD	81.69	35.05	58.37	44.19	88.62	26.39	57.50	38.34	97.58	70.48	84.28	79.44
GeSCF†	81.63	26.16	53.89	41.76	90.39	16.23	53.31	27.04	96.63	58.81	77.72	75.39
TERDNet	93.56	55.30	74.43	65.81	93.17	35.07	64.12	46.24	98.02	73.80	85.91	83.43
Ours	**94.61**	**62.83**	**78.72**	**69.65**	95.02	**50.03**	**72.52**	**61.73**	**98.28**	**77.80**	**88.03**	**86.3**
Ours-foundation	93.69	58.90	76.34	66.92	**95.56**	46.33	70.95	54.97	97.92	73.33	85.63	81.61

session comparison and the GFM's implicit 3D prior, Argos better distinguishes true scene changes from viewpoint-induced appearance differences. Qualitative results are shown in Appendix A.7.

5.4 Scene Change Detection (SCD) for Paired-image

Table 4 shows that, with only one image per session, Argos achieves the highest change IoU and F1 on PSCD, ChangeSim, and VL-CMU-CD. These results indicate that Argos's advantage is not limited to additional video context: the GFM's implicit 3D features provide stronger change descriptors than conventional 2D features. The same architecture therefore delivers strong performance in both video and traditional paired-image change detection.

5.5 Ablation Studies

Table 5: **Architecture Ablation.** Inter-session cross attention and Camera reader complement each other to achieve the best performance, while VGGT-Ω backbone provides a strong foundation.

Architecture Variant				VSCD**				PSCD			
Backbone	CDHead	CrossAttention	CameraReader	Static	Change	mIoU	F1	Static	Change	mIoU	F1
SAM	✓	✓	✓	91.92	27.89	59.90	30.64	92.05	46.08	69.07	58.59
VGGT-Ω	✓	×	×	96.34	47.97	72.16	42.52	94.64	63.88	79.27	69.26
VGGT-Ω	✓	✓	×	97.89	63.51	80.70	50.01	94.48	62.11	78.29	69.08
VGGT-Ω	✓	×	✓	93.74	25.96	59.85	29.37	94.63	63.53	79.08	69.16
VGGT-Ω	✓	✓	✓	**98.24**	**68.69**	**83.46**	**53.14**	**94.80**	**64.56**	**79.68**	**69.89**

Table 5 ablates our architectural components. The Change Detection Head on GFM features already outperforms the baselines, while replacing the backbone with SAM features degrades performance, supporting the value of the GFM's implicit 3D prior. On VSCD, Inter-session Cross-Attention further improves performance, with the Camera Reader providing additional gains when combined

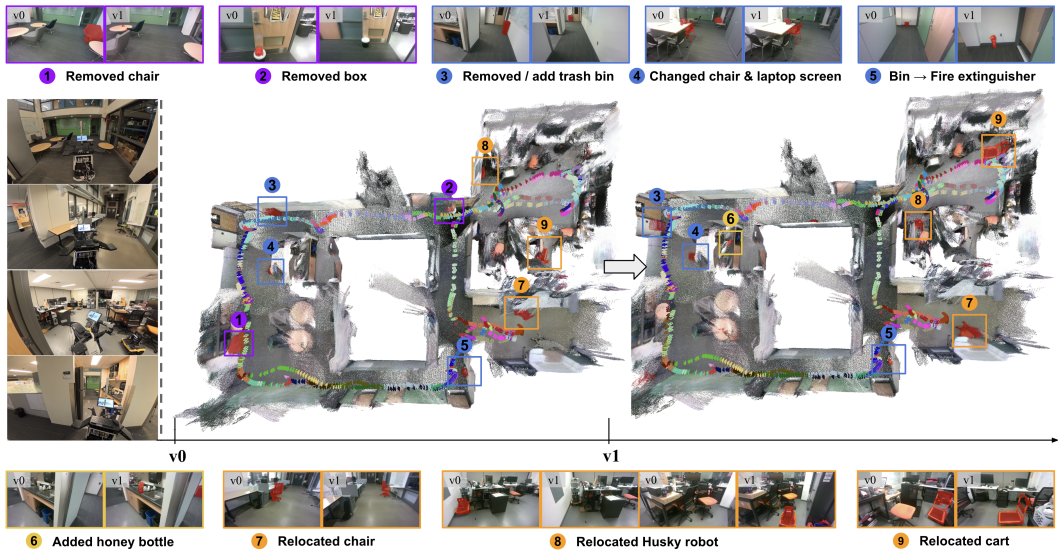


Figure 5: **Qualitative result of the online system in a multi-room scene.** Argos-SLAM localizes additions and removals across object scales, such as a removed chair, an added honey bottle, and a relocated Husky robot, consistently in both 2D observations and the global 4D map.

with it despite hurting performance in isolation. This suggests that pose information is most effective when used to guide cross-session comparison. The benefit is larger on video-based VSCD, where multi-view co-visibility helps resolve viewpoint and occlusion ambiguities. Additional analysis is provided in Appendix A.5.

5.6 Real-world Experiment

To evaluate Argos-SLAM in realistic settings, we conduct offline human-captured and online robotic experiments across small, medium, and multi-room indoor scenes with 4–10 object changes. Human trajectories use unposed iPhone videos, while robotic deployments use an Agilex mobile manipulator with an Intel RealSense D455. Both offline and online inference run on an NVIDIA RTX 5090 desktop GPU; during online deployment, the robot streams camera images to the workstation via ROS 2. As shown in Figure 5, Argos-SLAM operates under clutter, occlusion, and viewpoint changes, consistently localizing additions and removals across object scales in both image space and the reconstructed change-aware 4D map. With inputs capped at 64 images, each change-detection forward pass takes only 0.59 s. During online deployment, the SLAM loop produces a new submap every 8.2 s on average, while the corresponding asynchronous change-detection job completes in 6.9 s, allowing change updates to keep pace with reconstruction and typically finish before the next submap is produced. The accompanying video demonstrates the full on-the-fly reconstruction and map-update process; additional results on other scene scales are provided in Appendix A.6.

6 Conclusion

Detecting environmental changes and reconstructing consistent spatio-temporal maps remain critical yet challenging tasks for long-term robot autonomy. In this paper, we introduced Argos, which leverages implicit 3D knowledge from GFMs and a co-visibility-aware spatio-temporal comparator for joint scene change detection and 3D reconstruction. Training across diverse datasets, including our new benchmark, enables strong generalization across learned and zero-shot settings and takes a step toward a foundation model for scene change detection. We further introduced Argos-SLAM for online change-aware 4D mapping over long sequences, bridging change detection and long-term robot mapping. More broadly, our results demonstrate the potential of implicit geometric knowledge for change detection and encourage further research on other spatial reasoning tasks.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for implementing methods, cleaning and reformatting the dataset, providing feedback on research methodology or experiments, and supporting qualitative and thematic data analysis. We have not used generative AI tools to generate synthetic data sets, interpret results, propose or refine hypotheses, help develop theoretical models or conceptual frameworks, and design research methodology or experiments, and formulate mathematical claims, provide critical ingredients for proving mathematical claims, and assist in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, summarize or analyse existing literature, draft parts of a research paper, brainstorm, sourcing/searching for information, edit a research paper to improve readability, format references, and discuss word choice for the title for a research paper. We have reviewed all AI-assisted work. LLM generated code are verified by 2 authors, and evaluation results generated are compared with open-source implementations to verify evaluation correctness. LLM implementation of the proposed method is read carefully to ensure the code aligns with the implementation. LLM cleaned and reformatted datasets are being visually verified by 2 authors. LLM feedback on methods and experiments is discussed among the group and ensures a proper understanding without blind trust. We write the paper with an LLM, but we write the text initially by hand and let the LLM correct the grammar. The generated text was read carefully to ensure correctness and readability. We use an LLM to search for relevant papers to understand the topic more deeply. The retrieved papers are read carefully to ensure understanding of the topic. We brainstormed research ideas with LLMs, but none of the ideas were taken. In fact, this paper idea was not produced by an LLM. We did not use LLM-generated design for the model architecture, and all architecture choices are proposed by the authors after intuitive understanding of different architecture choices and the backbone's behavior. We used an LLM to correct the references. Specifically, the LLM helped to correct the conference name, and those name are verified by 1 author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We state that the work done in this paper is reproducible. In the appendix and the method section, we describe the details of the parameter choice for the model architecture, loss function, training dataset, training strategy, and training setup that help the reader to reproduce the model we use. The backbone model implementation is based on a public code base, and it is easy to adapt our method for implementation. We also describe the details about the evaluation protocol in the main text and the appendix, including what modifications to the dataset we made and what settings each dataset has, for readers to reproduce the results shown in the main text. For the proposed new dataset, we also illustrate the step-by-step generation details with the parameters we use to generate the results. The datasets are based on publicly available simulation assets and an annotation pipeline; readers can easily refer to the code base along with our description to generate the data. The real-world experiment methods and settings are described in detail in the main text and appendix, including the hardware type, software packages, and the method implementation details. We will also release the dataset and code in a proper open-source package to provide an official implementation for reference.

REFERENCES

Pablo F. Alcantarilla, Simon Stent, Germán Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42(7):1301–1322, October 2018. ISSN 1573-7527. doi: 10.1007/s10514-018-9734-5. URL https://doi.org/10.1007/s10514-018-9734-5.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3d reconstruction models, 2026. URL https://arxiv.org/abs/2512.11508.

C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J.J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robotics*, 32(6):1309–1332, 2016. ISSN 1552-3098. doi: 10.1109/TRO.2016.2624754.

Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *IEEE International Conference on Image Processing (ICIP)*, pp. 4063–4067, 2018. doi: 10.1109/ICIP.2018.8451652.

Shuo Chen, Kailun Yang, and Rainer Stiefelhagen. Dr-tanet: Dynamic receptive temporal attention network for street scene change detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pp. 502–509, 2021. doi: 10.1109/IV48863.2021.9575362.

Kyusik Cho, Dong Yeop Kim, and Euntai Kim. Zero-shot scene change detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 2509–2517, 2025.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 303–312, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237269. URL https://doi.org/10.1145/237170.237269.

Marius Fehr, Fadri Furrer, Ivan Dryanovski, Jürgen Sturm, Igor Gilitschenski, Roland Siegwart, and Cesar Cadena. Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5237–5244, 2017. doi: 10.1109/ICRA.2017.7989614.

Jiahui Fu, Chengyuan Lin, Yuichi Taguchi, Andrea Cohen, Yifu Zhang, Stephen Mylabathula, and John J. Leonard. Planesdf-based change detection for long-term dense mapping. *IEEE Robotics and Automation Letters*, 7(4):9667–9674, 2022. doi: 10.1109/LRA.2022.3191794.

Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald Dansereau, Niko Sunderhauf, and Dimity Miller. Multi-view pose-agnostic change localization with zero labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11600–11610, 2025.

Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald Dansereau, Niko Sunderhauf, and Dimity Miller. Changes in real time: Online scene change detection with multi-view fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Hyeonjae Gil, Dongjae Lee, Giseop Kim, and Ayoung Kim. Ephemerality meets lidar-based lifelong mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3312–3319, 2025. doi: 10.1109/ICRA55743.2025.11127618.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, second edition, 2003.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

Seoyeon Jang, Alex Junho Lee, I Made Aswin Nahrendra, and Hyun Myung. Chamelion: Reliable change detection for long-term lidar mapping in transient environments. *IEEE Robotics and Automation Letters*, 11(4):4361–4368, 2026. doi: 10.1109/LRA.2026.3665079.

Shyam Sundar Kannan and Byung-Cheol Min. Zeroscd: Zero-shot street scene change detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4665–4671, 2025. doi: 10.1109/ICRA55743.2025.11128082.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16384–16393, 2024. URL https://doi.org/10.1109/CVPR52733.2024.01550.

Jae-Woo Kim and Ue-Hwan Kim. Towards generalizable scene change detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24463–24473, June 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Chun-Jung Lin, Sourav Garg, Tat-Jun Chin, and Feras Dayoub. Robust scene change detection using visual foundation models and cross-attention mechanisms. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8337–8343, 2025. doi: 10.1109/ICRA55743.2025.11128568.

D. Maggio and L. Carlone. VGGT-SLAM 2.0: Real-time dense feed-forward scene reconstruction. In *Robotics: Science and Systems (RSS)*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Rafael Padilla, Allan F da Silva, Eduardo AB da Silva, and Sergio L Netto. Change detection in moving-camera videos with limited samples using twin-cnn features and learnable morphological operations. *Signal Processing: Image Communication*, 115:116969, 2023.

Jin-Man Park, Jae-Hyuk Jang, Sahng-Min Yoo, Sun-Kyung Lee, Ue-Hwan Kim, and Jong-Hwan Kim. Changesim: Towards end-to-end online scene change detection in industrial indoor environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8578–8585. IEEE Press, 2021. doi: 10.1109/IROS51168.2021.9636350. URL https://doi.org/10.1109/IROS51168.2021.9636350.

Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. Scenesmith: Agentic generation of simulation-ready indoor scenes. In *International Conference on Machine Learning (ICML)*, 2026.

R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 28085–28128, 2025.

Ragav Sachdeva and Andrew Zisserman. The change you want to see (now in 3d). In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Ragav Sachdeva and Andrew Zisserman. The change you want to see. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023b.

Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. In *British Machine Vision Conference (BMVC)*, 2015. URL https://api.semanticscholar.org/CorpusID:5634885.

Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6861–6867, 2020. doi: 10.1109/ICRA40945.2020.9196985.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9339–9347, 2019.

L. Schmid, M. Abate, Y. Chang, and L. Carlone. Khronos: A unified approach for spatio-temporal metric-semantic SLAM in dynamic environments. In *Robotics: Science and Systems (RSS)*, 2024.

Lukas Schmid, Jeffrey Delmerico, Johannes L. Schönberger, Juan Nieto, Marc Pollefeys, Roland Siegwart, and Cesar Cadena. Panoptic multi-tsdfs: a flexible representation for online multi-resolution volumetric mapping and long-term dynamic scene consistency. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8018–8024, 2022. doi: 10.1109/ICRA46639.2022.9811877.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Luiz GC Tavares, Allan F Da Silva, Rafael Padilla, Lucas A Thomaz, Sergio L Netto, and Eduardo AB Da Silva. Pixel-based change detection in moving-camera videos using twin convolutional features on a data-constrained scenario. *IEEE Access*, 13:123613–123629, 2025.

Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P. Balamuralidhar. Changenet: A deep learning architecture for visual change detection. In *European Conference on Computer Vision (ECCV) Workshops*, September 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7657–7666, 2019. URL https://api.semanticscholar.org/CorpusID:201070582.

Guo-Hua Wang, Bin-Bin Gao, and Chengjie Wang. How to Reduce Change Detection to Semantic Segmentation. *Pattern Recognition*, 138:109384, June 2023. ISSN 00313203. doi: 10.1016/j.patcog.2023.109384. URL https://linkinghub.elsevier.com/retrieve/pii/S0031320323000857.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yuqun Wu, Chih-hao Lin, Henry Che, Aditi Tiwari, Chuhang Zou, Shenlong Wang, and Derek Hoiem. Scenediff: A benchmark and method for multiview object change detection. *arXiv preprint arXiv:2512.16908*, 2025.

Jiae Yoon and Ue-Hwan Kim. VSCD: Video-based scene change detection in unaligned scenes. In *International Conference on Machine Learning (ICML)*, 2026a. URL https://openreview.net/forum?id=jjYEPizDXZ.

Jiae Yoon and Ue-Hwan Kim. Terdnet: Transformer encoder-recurrent decoder network for scene change detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, June 2026b.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one d4rt at a time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Hao Zhang, Mohamed El Banani, Jen-Hao Cheng, Paul Zhang, Yi Hua, Ben Mildenhall, Christoph Lassner, Narendra Ahuja, and Gengshan Yang. World tracing: Generative pixel-aligned geometry beyond the visible, 2026b. URL https://arxiv.org/abs/2606.13652.

Shibo Zhao, Guofei Chen, Honghao Zhu, Zhiheng Li, Changwei Yao, Nader Zantout, Seungchan Kim, Wenshan Wang, Ji Zhang, and Sebastian Scherer. Supermap: A spatio-temporal slam system for visual-language navigation. In *Robotics: Science and Systems (RSS)*, 2026.

Kaichen Zhou, Yuhan Wang, Grace Chen, Gaspard Beaudouin, Fangneng Zhan, Paul Liang, and Mengyu Wang. Page-4d: Disentangled pose and geometry estimation for vggt-4d perception. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 36401–36414, 2026.

A Appendix

A.1 Model Implementation Details

Camera Reader. Each MLP_ℓ has two linear layers with a GELU activation between them, mapping width $2d$ to $d/2$ and back to d ($4096 \to 1024 \to 2048$). Only the patch tokens are updated; the camera and register tokens of each layer pass through unchanged.

Inter-session Cross-Attention. Each $\mathrm{CrossAttn}_\ell$ uses 16 heads ($d_h = d/16$) with pre-normalization, applying separate layer normalizations to the queries and to the keys/values before the learned projections. The tokens of all frames in a visit are flattened into a single sequence, so each token attends to every token of every frame in the other visit, and both directions read the original features $Z^\ell_{v_0}$ and $Z^\ell_{v_1}$. The feed-forward sublayer is

$$\tilde{Z}^\ell_v = \bar{Z}^\ell_v + \mathrm{FFN}_\ell\big(\mathrm{LN}_\ell(\bar{Z}^\ell_v)\big), \qquad v \in \{v_0, v_1\}, \tag{6}$$

where LN_ℓ denotes layer normalization and FFN_ℓ is a two-layer network with GELU activation and hidden width $4d$.

Loss Function. The weighted binary cross-entropy is computed over the valid pixels of all images in a batch,

$$\mathcal{L} \;=\; \frac{1}{\sum_r |\Omega_r|} \sum_r \sum_{u \in \Omega_r} \mathrm{BCE}_{w^+}\big(z_r(u),\, M_r(u)\big), \tag{7}$$

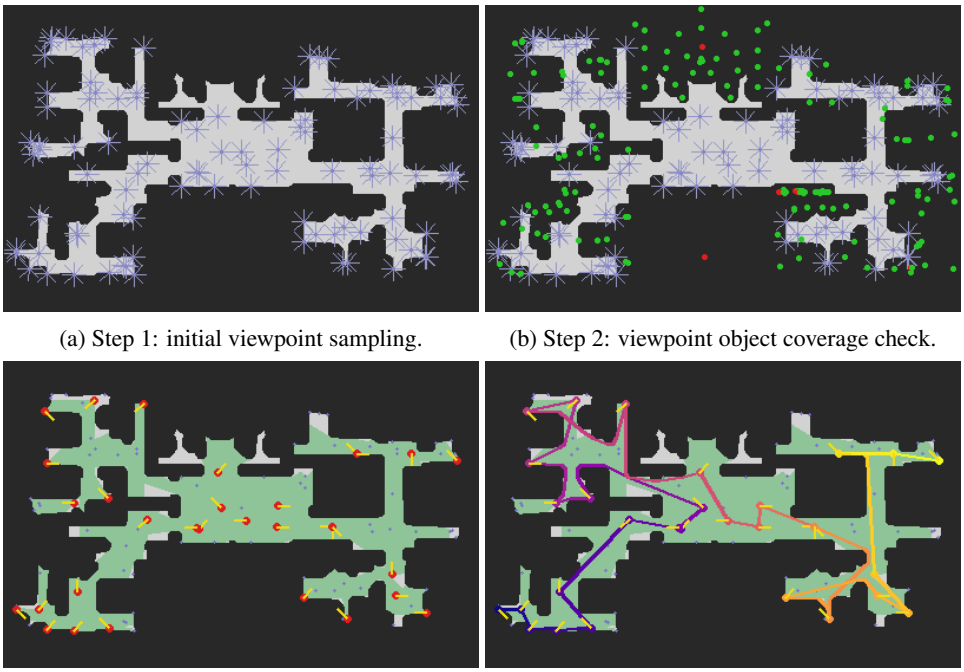
where $\Omega_r \subseteq \{1, \dots, H\} \times \{1, \dots, W\}$ is the set of valid pixels of image r, M_r is its ground-truth mask, and the positive (change) class is up-weighted by $w^+ = 5$. Frames without a pixel-aligned label (e.g., the unlabeled image of an SCD pair annotated in only one visit) receive $\Omega_r = \emptyset$ and are excluded from the loss.

Training Details. The Change Detection Head is warm-started from the pretrained depth head, whose DPT decoder has the same architecture, and only its final output layers are re-initialized; the Camera Reader and Inter-session Cross-Attention are randomly initialized. Each batch draws its number of images from $\{2, 8, 16, 32, 48, 64\}$ and an aspect ratio from $[0.5, 2.0]$, with the batch size chosen so that each GPU processes at most 64 images at resolution 256. The paired-image and video streams are served by separate loaders and interleaved by deficit scheduling: each step draws from the stream that has consumed the smaller fraction of its epoch. This spreads both data types evenly over training, giving a balanced learning signal throughout rather than long runs of a single type, and lets both streams finish together despite variable video batch sizes. For augmentation, we apply color jitter, a horizontal flip shared by all frames of a sample, and a random swap of the visit order with probability 0.5, so the shared cross-attention weights see both directions equally. We use AdamW with weight decay 0.05 and a peak learning rate of 10^{-5}, with linear warm-up over the first 5% of steps followed by cosine decay, gradient clipping at norm 1.0, and bfloat16 mixed precision. We train for 50 epochs with DDP on two RTX 4090 GPUs, which takes about three days.

A.2 Dataset Generation

Synthetic Dataset Generation. We build the synthetic portion of our dataset in five steps:

1. *State sampling.* For each scene, we generate three states by randomly removing movable objects. A support graph inferred from object geometry preserves physical plausibility by removing dependent objects together with their supports, while structural elements required for navigation are retained.
2. *Coverage planning.* Candidate viewpoints are sampled over the navigable space and evaluated for object visibility. We select viewpoints that jointly cover scene objects and navigable areas, then connect them with an open-path TSP to produce an independent trajectory for each state. The pipeline is illustrated in Figure 6.
3. *Rendering.* Each state is rendered along its trajectory at 512×512 and 30 fps, with camera orientation varying independently of travel direction to increase viewpoint diversity.
4. *Change masks.* Ground-truth masks are derived from object identities rather than pixel differences. For each ordered state pair, objects present in one state but absent from the



(a) Step 1: initial viewpoint sampling.

(b) Step 2: viewpoint object coverage check.

(c) Step 3: fog-of-war viewpoints selection.

(d) Step 4: TSP routing and trajectory generation.

Figure 6: **Synthetic dataset trajectory generation process.** The steps ensure maximum coverage of visible objects with a small number of viewpoints to ensure sufficient and efficient coverage of the environments.

other are marked as changed and rendered with a dedicated semantic label, yielding exact per-frame masks.

5. *Cross-visit matching.* Because visits follow independently generated trajectories, we match frames across states using scene overlap and pose similarity, retaining the best cross-visit correspondences for training and evaluation. Figure 7 shows that matched views share substantial visible scene geometry.

Argos-CD-HSSD and Argos-CD-SceneSmith share this generation pipeline, differing mainly in scene source and simulator. Argos-CD-HSSD uses HSSD scenes (Khanna et al., 2024) rendered in Habitat (Savva et al., 2019; Szot et al., 2021). Argos-CD-SceneSmith applies the same pipeline to procedurally generated SceneSmith scenes (Pfaff et al., 2026), using Drake's VTK renderer and a walkable navmesh derived from scene geometry. We provide both whole-house and per-room variants.



Figure 7: **Frustum overlap view match.**

Real-World Dataset Generation. We collect Argos-CD-real from independently captured revisits of real indoor scenes. Changes are annotated using the SceneDiff tool (Wu et al., 2025): annotators identify changed objects and provide sparse point prompts, which SAM2 propagates across both videos to produce dense per-frame masks. This supports real-world additions, removals, modifications, and deformable changes. We refer readers to the original paper for further details on the annotation procedure.

Dataset Statistics Comparison Protocol. Table 1 compares our datasets with prior scene change detection benchmarks. *Places* approximates the number of visually distinct locations visited within

a trajectory. We cluster per-frame DINOv2–SALAD descriptors using a cosine-distance threshold of 0.7 and merge revisits to the same location. *Length* denotes the duration of a single visit, computed from its frame count at 30 fps. Both are reported as mean/minimum/maximum over visits. *Pairs* counts cross-visit comparisons: one pair corresponds to two images for SCD datasets or two video clips for VSCD datasets. *Views* counts frames with ground-truth change masks.

A.3 Online System Implementation Details

We retain prior visit v_0 and reconstruct current visit v_1 online, submap-by-submap, using VGGT-SLAM (Maggio & Carlone, 2026). Each submap triggers a bounded *change-detection (CD) job* that processes its keyframes with selected prior keyframes in one Argos forward pass and transfers predictions to sparse 3D points.

Cross-visit alignment. We reuse VGGT-SLAM's SALAD-based (Izquierdo & Civera, 2024) loop closure. Once a cross-visit closure aligns v_1 with the prior reconstruction from v_0, CD jobs can use geometrically comparable prior frames. Verified loop closures receive the highest pairing confidence, $c = 2$.

Coarse SALAD matching. Using cached SALAD descriptors, each current keyframe retrieves its nearest prior keyframe by L_2 distance d, accepting matches with $d < \delta = 0.9$. Matching is bidirectional, and accepted pairs receive $c = \max(0, 1 - d/\delta)$ before geometric verification.

Geometric reprojection gate. Candidate pairs are verified by unprojecting depth, reprojecting across views, and computing symmetric in-frame depth-consistent overlap. We reject overlap below 0.5, use a 15% depth tolerance, and assign accepted pairs their overlap score as $c \in [0, 1]$. To ensure, fast compute, verification uses an 8×8 strided pixel grid ($\sim 1/64$ of pixels); world points are computed once per frame and cached across comparisons.

Scene-centroid coverage search. For unmatched frames, we index prior keyframes by the centroid of their visible world points and geometrically evaluate only the $k = 10$ nearest centroids using the same reprojection score. This supports large viewpoint changes while reducing search from $\mathcal{O}(n_{v_0} n_{v_1})$ to $\mathcal{O}(n_{v_1} k)$.

Temporal fallback. Remaining unmatched frames inherit the correspondence of the nearest matched frame within $w = 5$ frames, with confidence $c = 0.1(1 - \Delta/(w + 1)) < 0.1$ for temporal distance Δ. Frames without a correspondence are excluded from change inference and cross-visit fusion.

Asynchronous CD execution. Matching and inference run on a background worker, keeping reconstruction responsive. Jobs respect Argos's 64-image limit by reducing the current-visit window while keeping matched pairs together.

Change-aware map fusion. Per-pixel change probabilities and pairing confidences are transferred to their corresponding sparse 3D points; for prior frames used in multiple jobs, we retain the highest confidence. After pose-graph optimization, observations within each world-space voxel are fused by a confidence-weighted mean. Pairing confidence measures correspondence quality and is distinct from the model's unused per-pixel confidence. For a queried visit v with other visit v', points from v' whose fused probability exceeds a threshold τ form the changed set $\mathbb{P}_{\Delta}^{v'}$; removing it from the joint reconstruction yields the change-aware point-cloud map $\mathbb{P}(v)$ of Section 4.5.

A.4 Evaluation Protocol

Adapting SCD Baselines to VSCD Benchmarks. Paired-image SCD methods assume correspondence between their two inputs. To evaluate them on video benchmarks, we establish cross-visit pairs using SALAD (Izquierdo & Civera, 2024): for each frame in one visit, we retrieve its best-matching frame from the other visit and apply the SCD model to the pair. We repeat this in both directions so that every frame in both visits receives a prediction.

Benchmark Setup. For PSCD and VL-CMU-CD, we follow the per-image protocols of Lin et al. (2025). For ChangeSim, we exclude the "removed" class from the ground-truth masks and report results under the original ChangeSim protocol in Table 8. For VSCD, we retain the original RGB videos but replace the original v_1-aligned labels with masks for both visits, where each mask marks

content present in that visit and absent from the other. We uniformly sample 32 frames per visit and report results under the original VSCD protocol separately in Table 9. For SceneDiff and Argos-CD-real, we sample 16 frames per visit. For the longer HSSD-CD and SceneSmith-CD sequences, we evaluate within fixed 48-frame temporal windows, sampling 16 frames from the current visit and selecting corresponding prior-visit frames using the cached cross-visit matches.

A.5 Additional Experiments

Table 6: **O-SCD pose-registration prescreen coverage.** Across all 220 classic-VSCD test candidates, increasing the reference-frame budget from 32 to 128 provide enough datapoint for reliable evaluation.

Setting	#Samples	COLMAP Success	Avg. Query Frames	Full Success (32/32)	≥90% Success (29+/32)
32+32	220	141 (64.1%)	7.5 / 32	4 (1.8%)	4 (1.8%)
64+32	220	181 (82.3%)	11.9 / 32	6 (2.7%)	15 (6.8%)
128+32	220	199 (90.5%)	13.8 / 32	7 (3.2%)	15 (6.8%)
256+32	220	195 (88.6%)	13.0 / 32 4	(1.8%)	9 (4.1%)

Table 7: **Comparison against O-SCD on the top-12 classic-VSCD scenes O-SCD can register.** When given enough reference frames, O-SCD outperforms the VSCDNet baseline but still falls short of our proposed method.

Method	Static	Change	mIoU	F1
O-SCD (R=32)	88.42	19.18	53.80	26.28
O-SCD (R=128)	90.20	<u>28.95</u>	<u>59.57</u>	<u>40.64</u>
VSCDNet	<u>90.22</u>	22.45	56.33	31.05
Ours (VSCD)	**93.21**	**37.70**	**65.46**	**49.41**

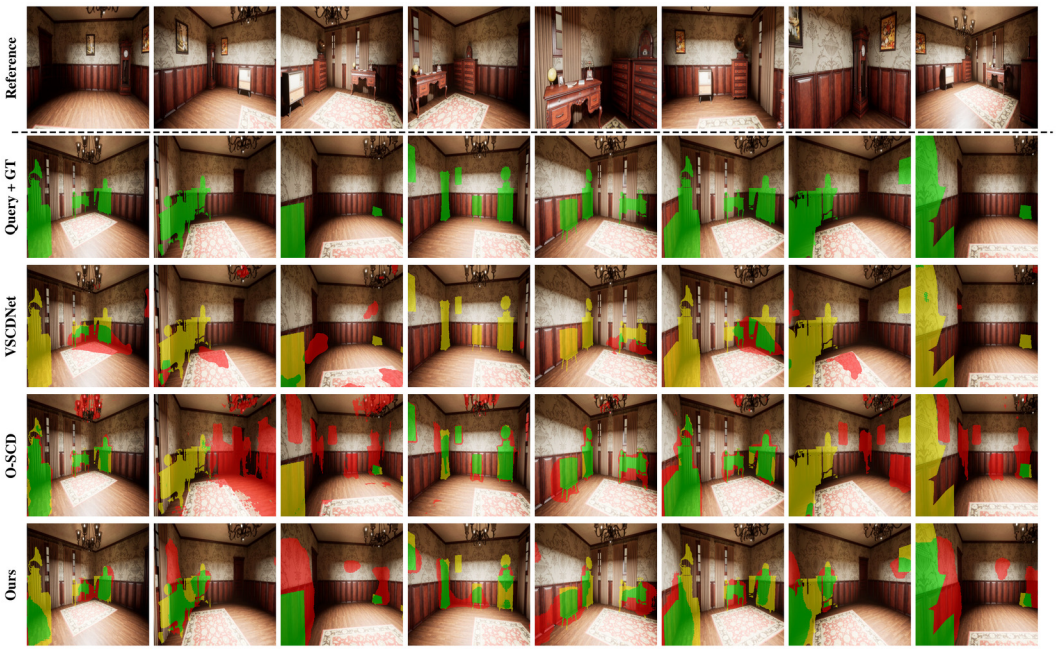
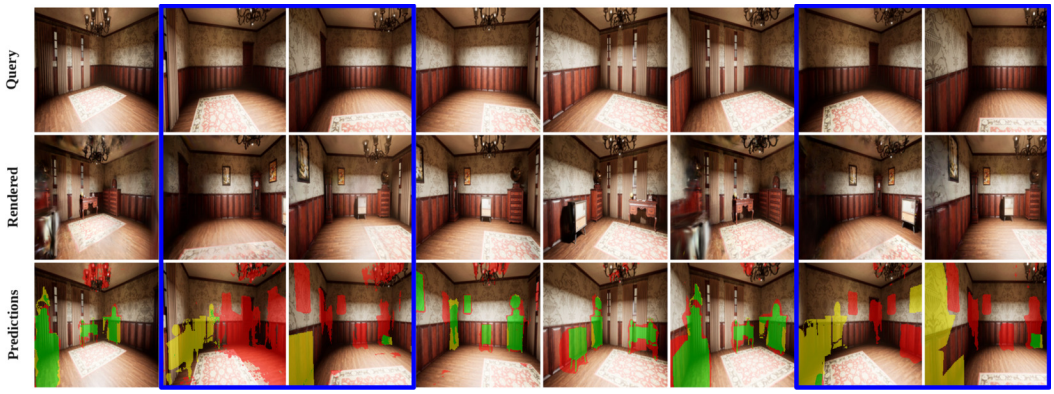


Figure 8: **Qualitative comparison between O-SCD, VSCDNet, and Argos.** Even with 128 reference frames, O-SCD produces sensible predictions—particularly over removed objects—but still yields many false positives (FP) and false negatives (FN).

Comparison against O-SCD. O-SCD predicts both "appeared" and "removed" masks for mask-only query frames given a set of reference frames, a formulation that differs from ours in Section 3;



(a) Poor rendering quality from the Gaussian Splats at novel views corresponding to the query camera positions.



(b) **Blue boxes** highlight incorrect pose estimation, which is readily identifiable from the carpet orientation.

Figure 9: **O-SCD limitation analysis.** Even when COLMAP and Gaussian Splatting succeed, poor rendering quality or incorrect camera registration can still lead to noisy predictions.

we therefore evaluate on the original VSCD dataset, which is consistent with O-SCD's output format. As shown in Table 6, under the original protocol of 32 reference and 32 query images, O-SCD frequently fails to produce results due to COLMAP or query-frame camera-pose registration failures. We thus provide 128 reference frames—a more generous setting that favors O-SCD—and select the top-12 samples on which it consistently succeeds. For this evaluation, we train our model on VSCD to predict both "appeared" and "removed" masks, and Table 7 shows that our method outperforms O-SCD even under settings that favor it. We further provide a qualitative comparison in Figure 8.

We further analyze O-SCD qualitatively (Figure 9) and identify two limitations that frequently hurt its pipeline: (1) poor rendering quality produced by noisy camera-position estimates or by query frames that fall outside the distribution of the reference frames; and (2) incorrect pose registration caused by large scene changes, e.g., when most objects are removed, which yields a well-rendered but wrong image for change detection.

Additional benchmark results. We additionally evaluate Argos under the original ChangeSim and VSCD protocols. For these experiments, Argos predicts both "appeared" and "removed" regions, which are combined into a binary change mask. For VSCD, we follow the original evaluation protocol and additionally compare against two abandoned-object detection (AOD) methods, TCF-LMO (Padilla et al., 2023) and PBCD-MC (Tavares et al., 2025).

Argos remains consistently stronger under these alternative output protocols. On ChangeSim (Table 8), it improves change IoU from 34.20 to 51.75 and F1 from 46.01 to 65.13 over the strongest baseline, TERDNet. On the original VSCD benchmark (Table 9), Argos outperforms prior methods on both synthetic and real-world data and remains strong across variations in video length, graphics

Table 8: **Performance on ChangeSim under original protocol.**

Method	ChangeSim			
	Static	Change	mIoU	F1
FC-EF	65.68	16.58	41.13	27.21
FC-Siam-diff	84.56	25.03	54.80	38.55
FC-Siam-conc	80.63	23.32	51.95	35.44
CSCDNet	90.61	24.05	57.33	35.70
DR-TANet	89.14	28.02	58.59	39.71
C-3PO	91.16	28.67	59.99	40.17
ZSSCD†	92.55	28.25	60.40	39.54
RobustSCD	87.12	28.85	57.82	41.41
GeSCF†	88.82	20.27	54.54	33.28
TERDNet	91.37	34.20	62.78	46.01
Ours	**93.87**	**51.75**	**72.81**	**65.13**

Table 9: **Overall performance on the original VSCD benchmark.** We report frame-wise F1 score (%) on the synthetic dataset and the real-world test set, together with results stratified by video length, graphic level, and the number of changed objects.

Method		Video length			Graphic level		Object changes			Synthetic	Captured by		Real-world
Task	Model	Low	Mid	High	Mid	High	Low	Mid	High	Overall	Robot	Human	Overall
AOD	TCF-LMO	17.1	20.1	22.2	18.7	20.8	15.2	21.0	21.8	19.7	7.4	12.0	10.3
	PBCD-MC	24.6	28.0	26.5	26.3	27.4	23.4	28.0	27.8	26.8	16.8	15.7	16.1
SCD	CSCDNet	18.8	21.4	16.9	21.7	17.5	20.0	20.1	19.0	19.8	8.2	9.6	9.1
	DR-TANet	20.5	22.0	17.2	22.9	17.9	20.8	20.5	20.6	20.6	9.1	13.1	11.6
	C-3PO	22.7	26.1	20.8	30.1	16.9	18.8	25.1	27.8	24.1	8.0	13.9	11.7
	ZSSCD	0.1	0.5	0.1	0.5	0.1	0.2	0.4	0.2	0.3	2.1	0.0	0.8
	GeSCF	26.4	30.4	31.2	29.3	29.8	28.0	29.7	30.6	29.5	25.1	12.6	17.3
VSCD	VSCDNet	38.1	36.9	33.9	40.7	31.7	32.1	37.7	39.0	36.6	**32.0**	21.5	25.4
	Ours*	**48.07**	**50.04**	**46.44**	**53.29**	**42.99**	**42.39**	**51.10**	**49.35**	**48.88**	26.90	**38.13**	**34.92**

quality, and the number of changed objects. These results suggest that Argos learns transferable change cues that remain effective across different annotation and output conventions.

Table 10: **Inference runtime vs. number of input images.**

Method	Number of input images N						
	2	4	8	16	32	64	128
TERDNet	0.076	0.095	0.13	0.26	0.51	1.03	2.05
C-3PO	0.008	0.010	0.015	0.026	0.049	0.10	0.21
VSCDNet	0.030	0.059	0.12	0.25	0.48	0.96	1.91
SceneDiff†	22.8	32.6	51.6	96.6	210	485	OOM
O-SCD‡	Fail	Fail	Fail	187	209	194	286
Ours-foundation	0.037	0.043	0.068	0.12	0.25	0.59	1.45

Runtime Comparison. We measure mean wall-clock time (seconds) of one forward pass (batch size 1, single RTX 5090) as the total number of input images N grows ($N/2$ reference and $N/2$ query frames); every method predicts a change mask for all N frames. † indicates zero-shot methods; SceneDiff's time includes loading its Pi3, SAM, and DINOv3 models due to multi-stage pipeline design and it runs out of host CPU memory (OOM, 125 GB RAM) at $N{=}128$. ‡ O-SCD reconstructs a 3D Gaussian scene from the reference session (via COLMAP structure-from-motion and 3DGS training) before online query localization, so its runtime is dominated by this offline step rather than the online inference itself; it fails (Fail) for $N{\leq}8$ because COLMAP cannot register a usable reconstruction from so few reference views.

A.6 Additional Real-World Qualitative Results

Figure 10 and Figure 11 show the mapping results from the Agilex mobile manipulator in small and medium environments. Figure 12 and Figure 13 show the offline mapping results from iPhone-captured data in a multi-room indoor apartment and a small office.

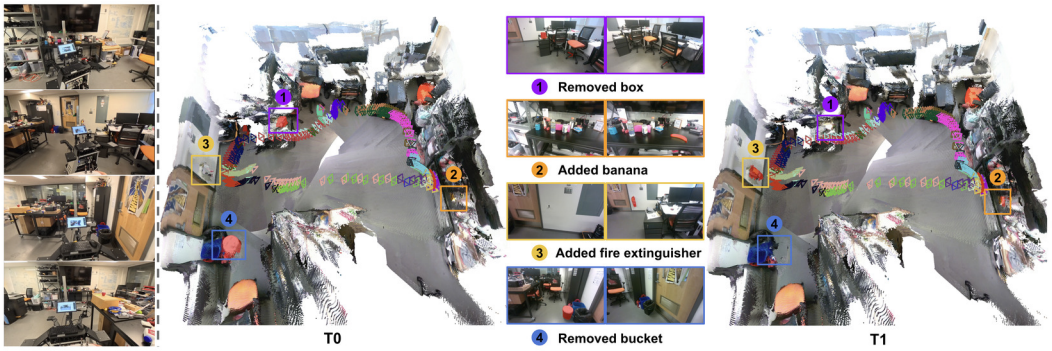


Figure 10: **Qualitative result of online robot mapping in a small-size workshop environment.**

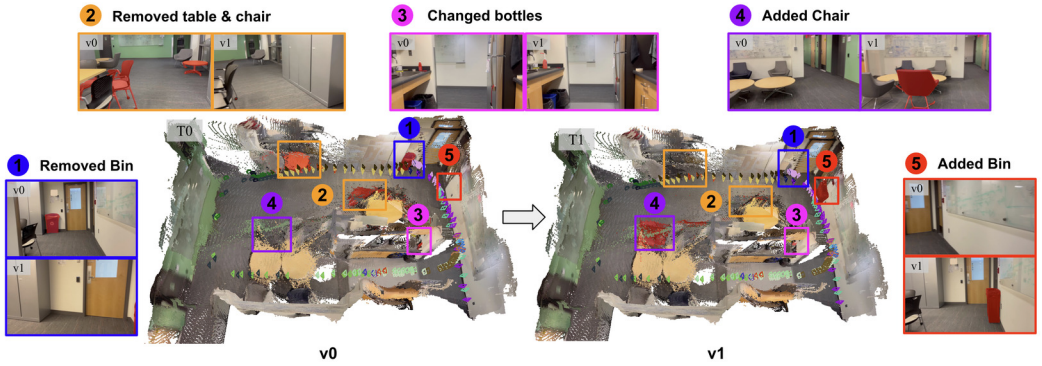


Figure 11: **Qualitative result of online robot mapping in a medium-sized common space area.**

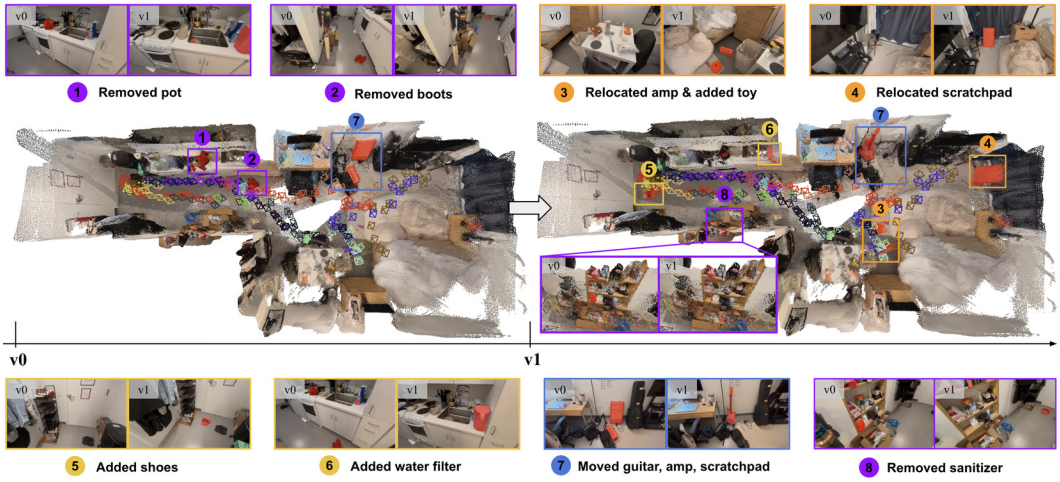


Figure 12: **Qualitative result of the offline mapping from iPhone-captured video in a multi-room apartment.**

A.7 Additional Qualitative Results From Change Detection Datasets

We provide additional qualitative results of sampled scenes from each dataset. SCD results are shown in Figure 14, and VSCD results are shown in Figure 15, 17, 16, 18, and 19.

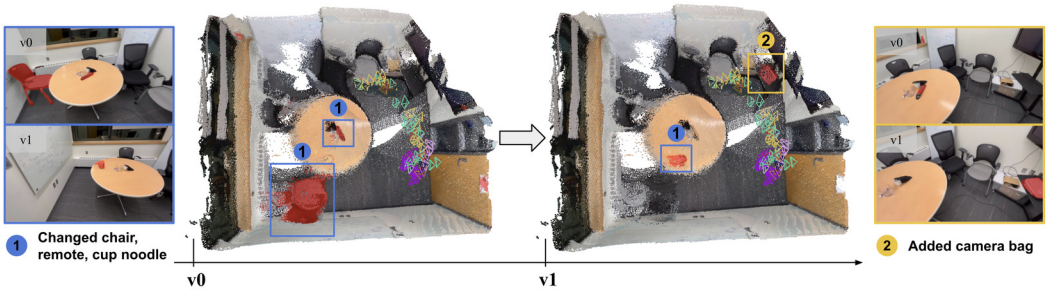
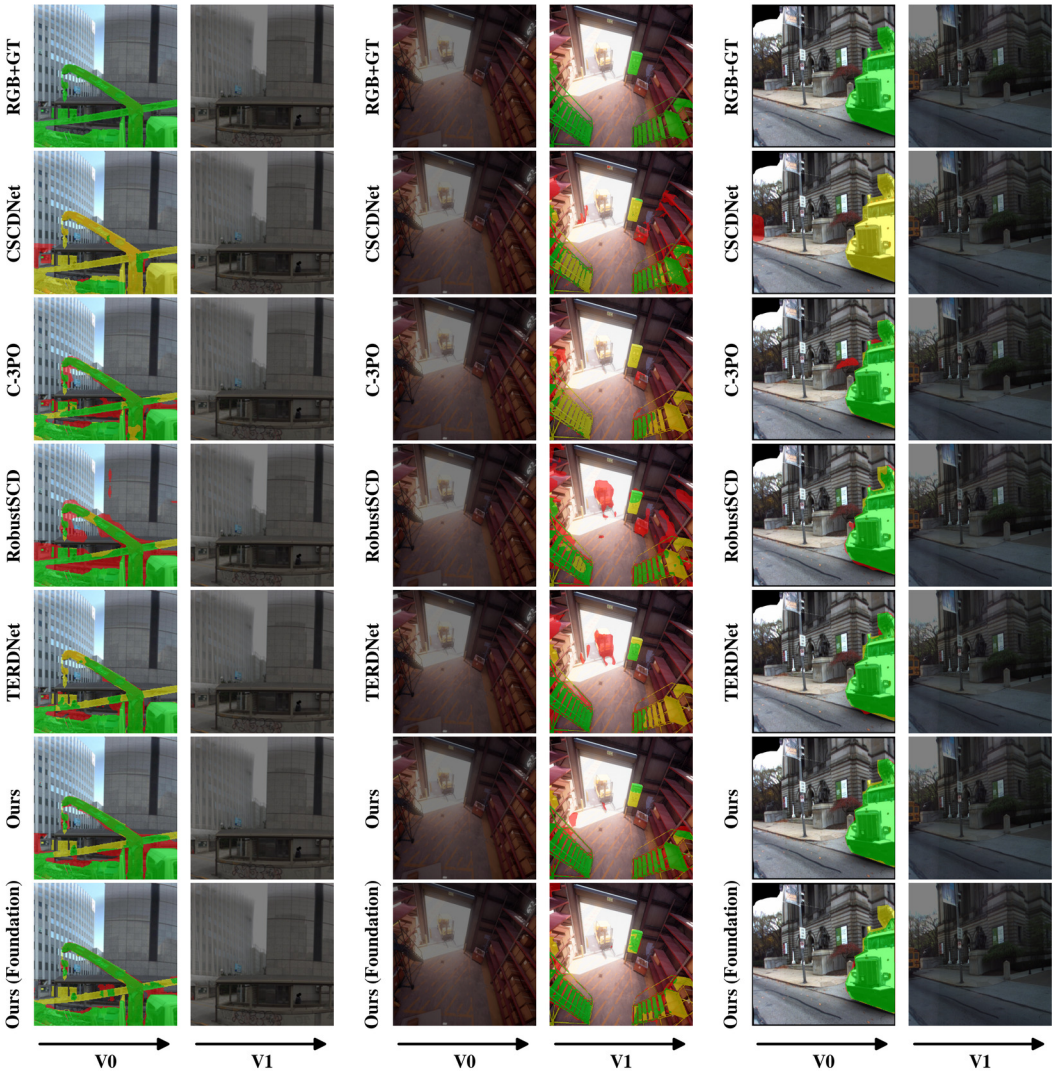


Figure 13: **Qualitative result of online robot mapping in a small-size office environment.**



(a) PSCD (b) ChangeSim (c) VL-CMU-CD

Figure 14: **Qualitative results of SCD datasets.** SCD contains image pairs with labels only on the v_0 or v_1 image. Argos consistently produces predictions with less noise.

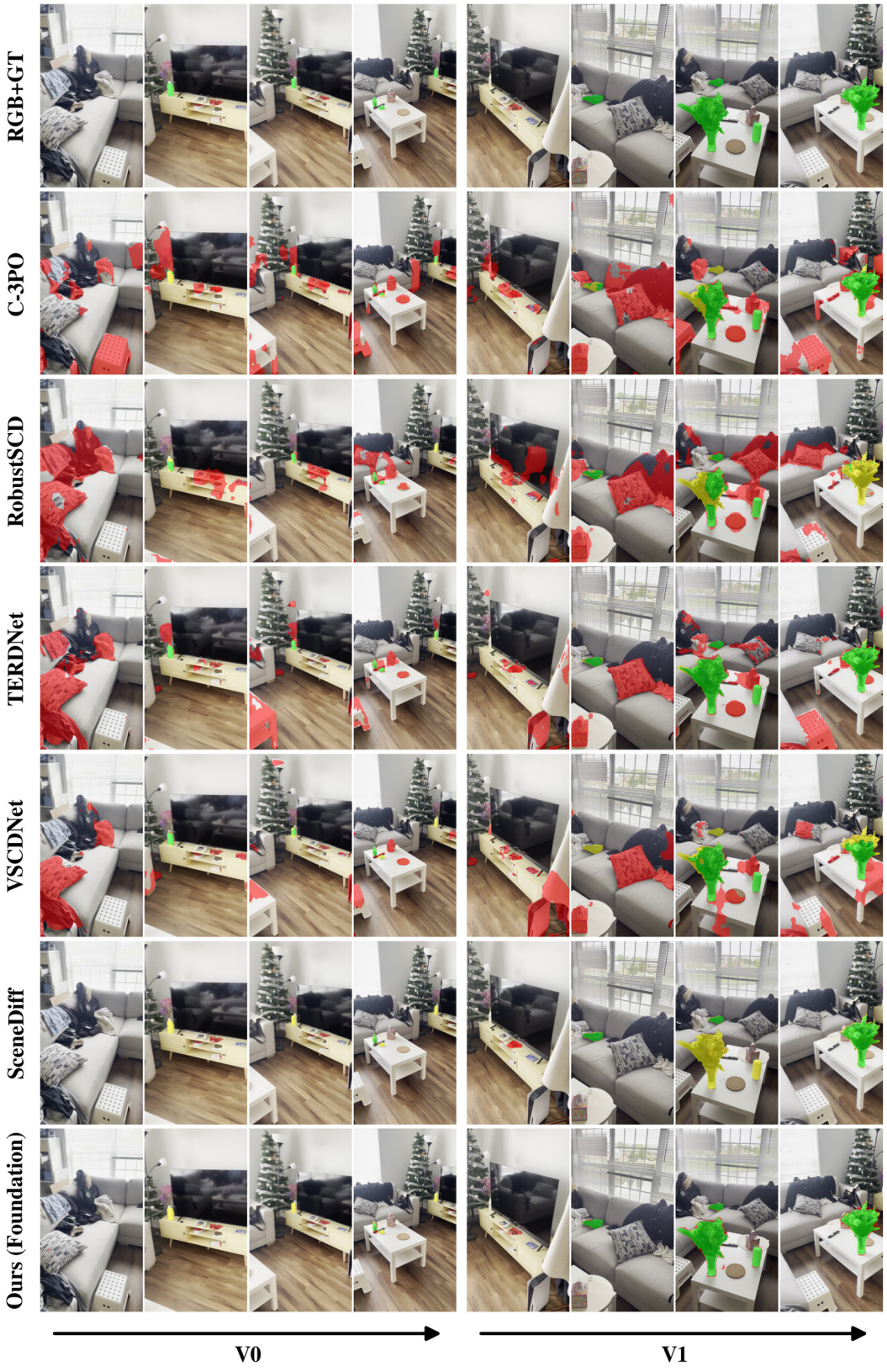


Figure 15: **Qualitative result of the SceneDiff dataset.**

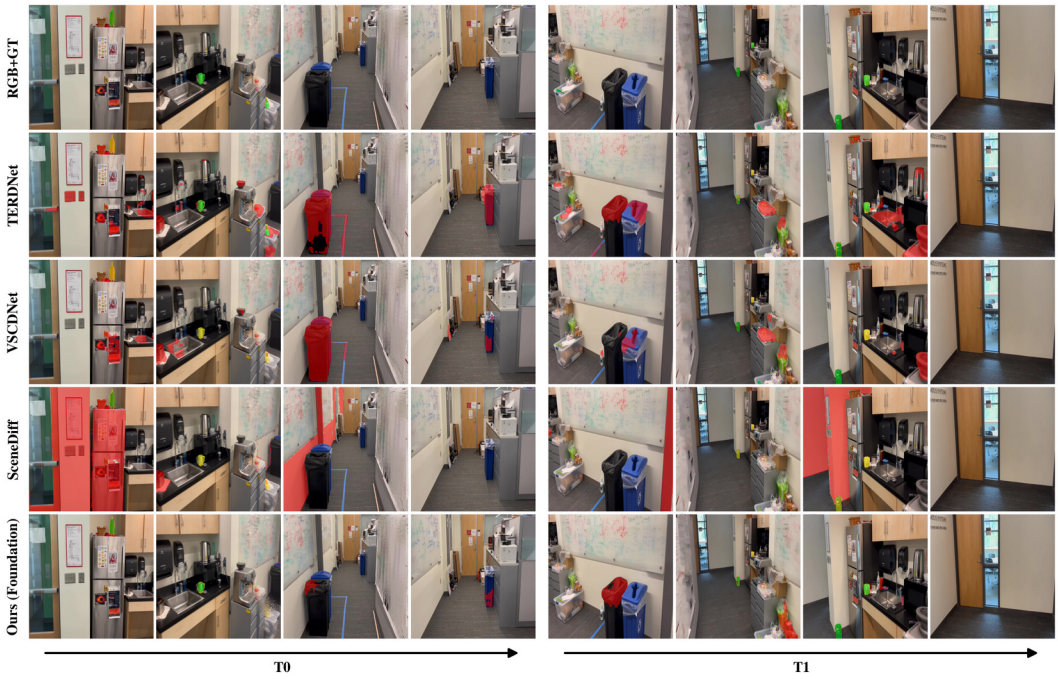


Figure 16: **Qualitative result of the Argos-CD-real dataset.** Stretched vertically to provide a better view.

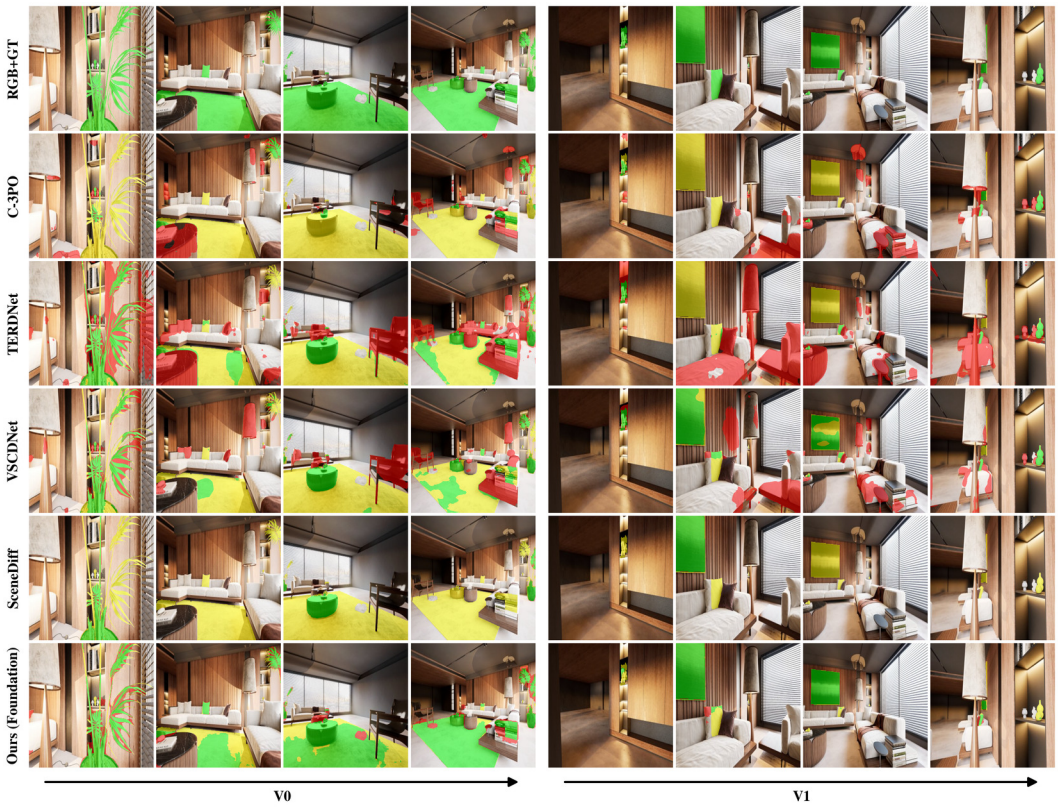


Figure 17: **Qualitative result of the VSCD dataset.** Stretched vertically to provide a better view.

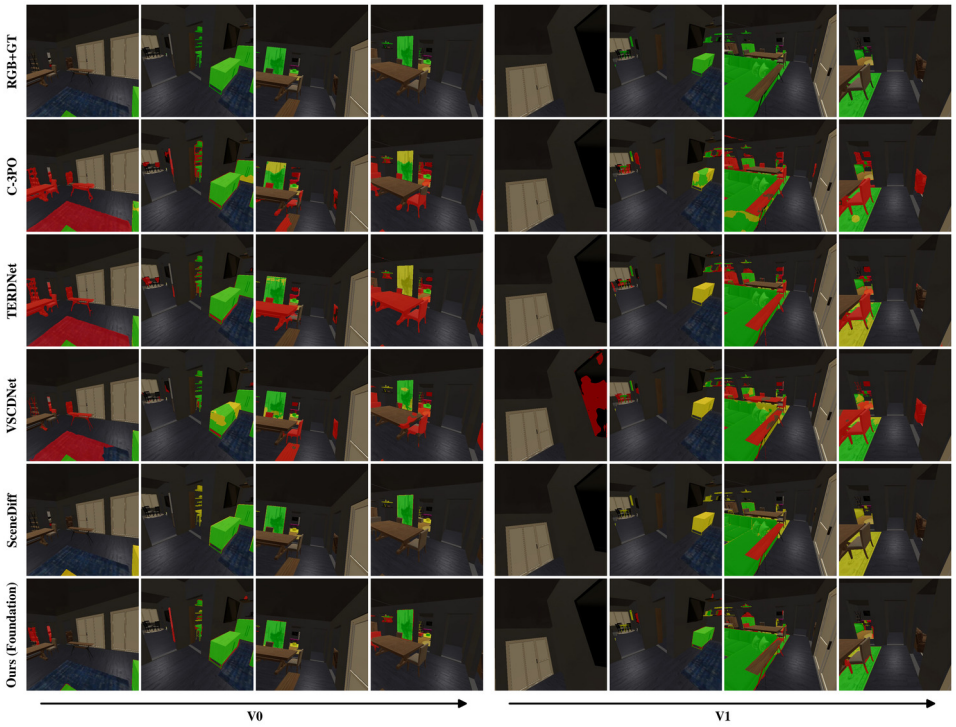


Figure 18: **Qualitative result of the Argos-CD-HSSD dataset.**

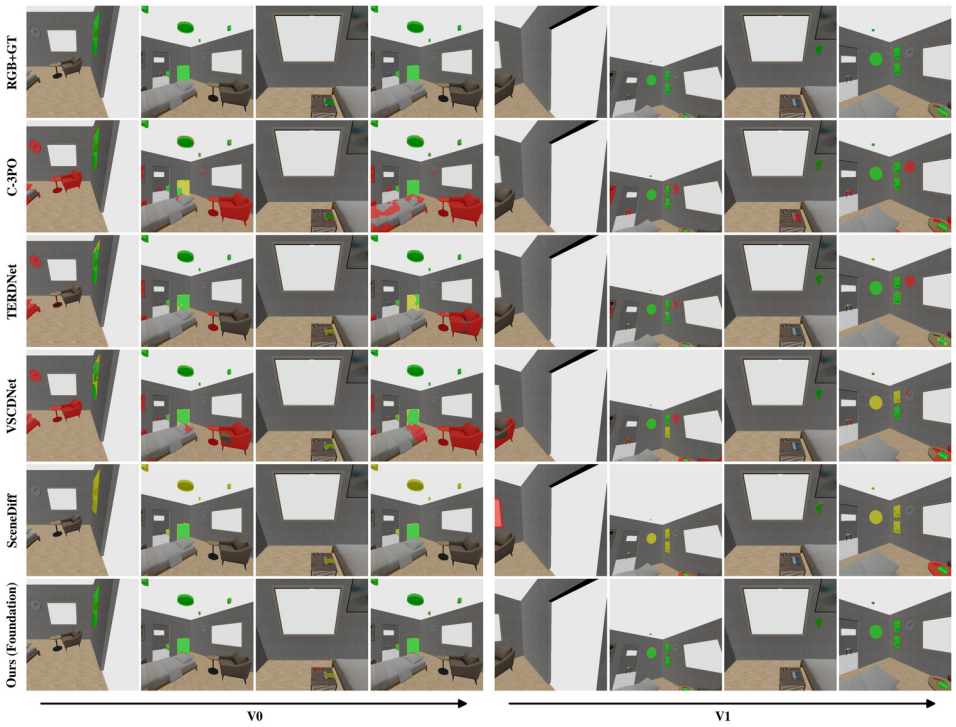


Figure 19: **Qualitative result of the Argos-CD-SceneSmith dataset.**